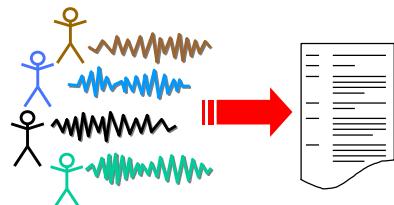


Rich Transcription 2003



Spring Speech-to-Text Evaluation Results

May 19, 2003

Overview

- Test Corpora
- Evaluation Conditions
- Scoring
- Results

Dimensions for Test Set / Cross-Year Comparisons

- Current vs. Progress Test Sets:
 - **Current Test** – a test based on fresh, recently-collected material. Permits system-system comparison, but cannot be used to directly track progress. This material is released as development data after the tests.
 - **Progress Test** (English only) – a test based on fixed, reusable material. Permits system-system and cross-test comparison. This material cannot be examined in any way and is not released after the test.
- Mothballed Systems (English Only)
 - Ran earlier systems (or comparable) on Progress Test Set to create 2002 baseline.
- These tools permit us to:
 - Calibrate progress by holding test set constant over time.
 - Calibrate test set difficulty by holding systems constant and running on different test sets.

RT-03S Test Corpora

Corpus	Reuse?	BN Scope	CTS Scope
		<i>TDT-4 Sources from Feb 2001, excerpts transcribed to nearest story boundary</i>	<i>5 min excerpts, two sides presented separately, transcribed to nearest turn</i>
Current English	No	180 minutes 6 shows, 1 per source 30-min. excerpts	360 minutes, 72 conv: 36 – unused SWBD Cell 36 – new Fisher
Current Chinese	No	60 minutes 5 shows, 1 per source 12-min. excerpts	60 minutes, 12 conv: 12 – unused CallFriend
Current Arabic	No	60 minutes 2 shows, 1 per source 30-minute excerpts	60 minutes, 12 conv: 12 – unused CallHome
Progress English	Yes	180 minutes 6 shows, 1 per source 30-min. excerpts	180 minutes, 36 conv: 36 – new Fisher

Test Data Selection

- **BN:**
 - Selected shows that cover the evaluation epoch months
- **CTS:**
 - Fisher:
 - Initial pool of 477 calls
 - Limited to speakers who participate in exactly one Fisher call
 - LDC performed QC on the calls
 - Selected two 36-call test sets (Current and Progress)
 - Balanced by gender, age, region, and land/cell 3:1 ratio
 - Excluded heavy accents, poor channel quality, non-conversational speech, buggy recordings
 - SWBD-Cell:
 - Initial pool limited to unused calls previously transcribed by the LDC
 - Selected one 36-call test set (Current)
 - Balanced by gender and device (land-land / land-cell / cell-cell)
 - Similar screening to above
 - CallHome and CalifFriend:
 - Same selection procedure as SWBD-Cell

STT Test Conditions

- Language
 - English, Chinese (Mandarin), Arabic (Egyptian)
- Domain
 - Broadcast News, Conversational Telephone Speech
- Processing time categories
 - 1X, 10X, Unlimited
- Progress vs. Current test sets
- “Mothballed” vs. State-of-the-art system runs

48 Supported STT Tests!

- Experiment ID system designed to track many experiments:
 - expt_03_stt10x_eval03_arab_bnews_spch_expt_1
 - expt_03_stt10x_eval03_arab_cts_spch_expt_1
 - expt_03_stt10x_eval03_eng_bnews_spch_expt_1
 - expt_03_stt10x_eval03_eng_cts_spch_expt_1
 - expt_03_stt10x_eval03_mand_bnews_spch_expt_1
 - expt_03_stt10x_eval03_mand_cts_spch_expt_1
 - expt_03_stt10xmb_eval03_arab_bnews_spch_expt_1
 - expt_03_stt10xmb_eval03_arab_cts_spch_expt_1
 - expt_03_stt10xmb_eval03_eng_bnews_spch_expt_1
 - expt_03_stt10xmb_eval03_eng_cts_spch_expt_1
 - expt_03_stt10xmb_eval03_mand_bnews_spch_expt_1
 - expt_03_stt10xmb_eval03_mand_cts_spch_expt_1
 - expt_03_stt11x_eval03_arab_bnews_spch_expt_1
 - expt_03_stt11x_eval03_arab_cts_spch_expt_1
 - expt_03_stt11x_eval03_eng_bnews_spch_expt_1
 - expt_03_stt11x_eval03_eng_cts_spch_expt_1
 - expt_03_stt11x_eval03_mand_bnews_spch_expt_1
 - expt_03_stt11x_eval03_mand_cts_spch_expt_1
 - expt_03_stt1xmb_eval03_arab_bnews_spch_expt_1
 - expt_03_stt1xmb_eval03_arab_cts_spch_expt_1
 - expt_03_stt1xmb_eval03_eng_bnews_spch_expt_1
 - expt_03_stt1xmb_eval03_eng_cts_spch_expt_1
 - expt_03_stt1xmb_eval03_mand_bnews_spch_expt_1
 - expt_03_stt1xmb_eval03_mand_cts_spch_expt_1
 - expt_03_sttul_eval03_arab_cts_spch_expt_1
 - expt_03_sttul_eval03_eng_bnews_spch_expt_1
 - expt_03_sttul_eval03_eng_cts_spch_expt_1
 - expt_03_sttul_eval03_mand_bnews_spch_expt_1
 - expt_03_sttul_eval03_mand_cts_spch_expt_1
 - expt_03_sttulmb_eval03_arab_bnews_spch_expt_1
 - expt_03_sttulmb_eval03_arab_cts_spch_expt_1
 - expt_03_sttulmb_eval03_eng_bnews_spch_expt_1
 - expt_03_sttulmb_eval03_eng_cts_spch_expt_1
 - expt_03_sttulmb_eval03_mand_bnews_spch_expt_1
 - expt_03_sttulmb_eval03_mand_cts_spch_expt_1
 - expt_03_stt10x_prog_eng_bnews_spch_expt_1
 - expt_03_stt10x_prog_eng_cts_spch_expt_1
 - expt_03_stt10xmb_prog_eng_bnews_spch_expt_1
 - expt_03_stt10xmb_prog_eng_cts_spch_expt_1
 - expt_03_stt11x_prog_eng_bnews_spch_expt_1
 - expt_03_stt11x_prog_eng_cts_spch_expt_1
 - expt_03_stt1xmb_prog_eng_bnews_spch_expt_1
 - expt_03_stt1xmb_prog_eng_cts_spch_expt_1
 - expt_03_sttul_prog_eng_bnews_spch_expt_1
 - expt_03_sttulmb_prog_eng_bnews_spch_expt_1
 - expt_03_sttulmb_prog_eng_cts_spch_expt_1

Processing Rules

- BN
 - Shows and broadcast dates are permitted side information
 - Test data must be processed in chronological order, look-ahead not permitted
 - Excerpts presented in separate files.
 - Test epoch: February 2001
 - No training/development data after test epoch begins
- CTS
 - No side information to be used
- For all conditions:
 - Speech Input only
 - Full waveforms distributed, but only specified excerpts could be processed.
 - No manual segmentation provided
 - Optional use of MIT-LL baseline segmentation
 - All processing must be automatic
 - automatic adaptation permitted

System Output

- Extended CTM
 - CTM + token type + (optional) speaker ID information
 - Token types:
 - Lexical token
 - Fragment
 - Filled pause
 - Miscellaneous – metadata to be passed through for MDE experiments
 - Token encoding
 - SNOR for English data
 - GB-encoded orthography for Mandarin data
 - Romanized orthography for Arabic BN data
 - UTF8-encoded orthography for Arabic CTS data
- System description for each run
 - Must include speed factor info
 - SF = total processing time/source file duration

STT Scoring

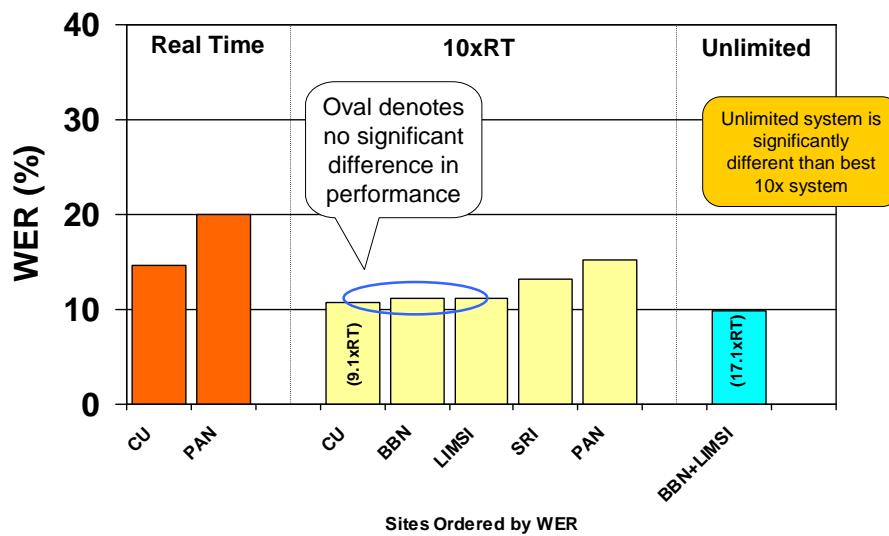
- Scored as in past using SCLITE
 - GLM updated with new contractions and compound words in test material
 - Overlapping speech excluded
- Only changes
 - Consistent orthography required for English data
 - Spelled letters must be of form “a.”, “b.”, “c.”, ...
 - No non-alphabetic characters except apostrophes and hyphens
 - Token type-based scoring
 - Tokens not of type “lex” are removed prior to scoring from hyp and ref
 - Old training data fixes removed from GLM
 - Did not have significant effect on dev test scores

BN STT Results

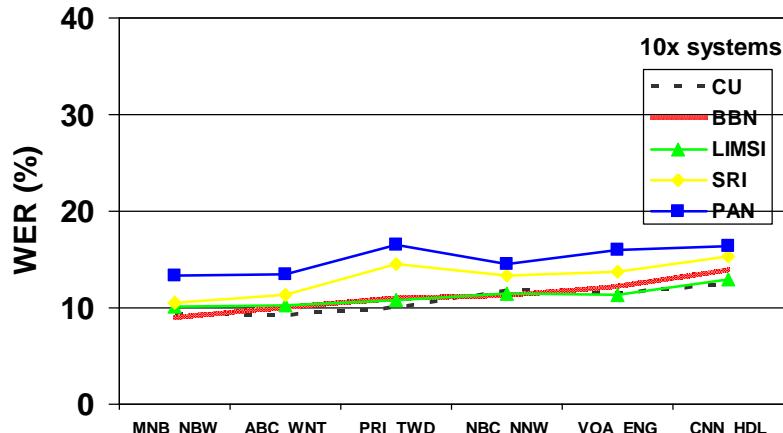
RT-03 Systems

English BN STT Results

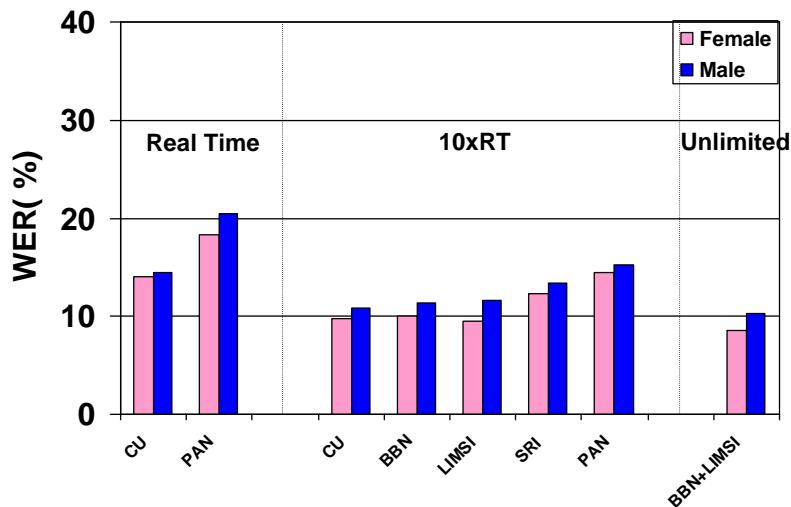
Current Test Set



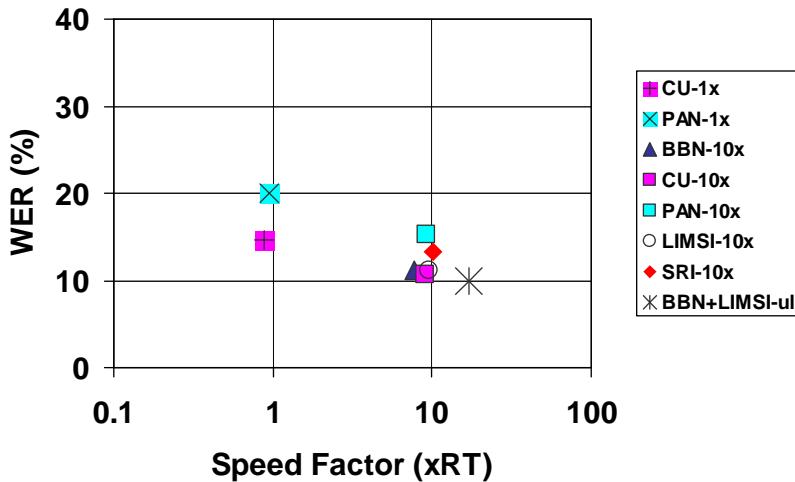
English BN WER by Show Current Test Set



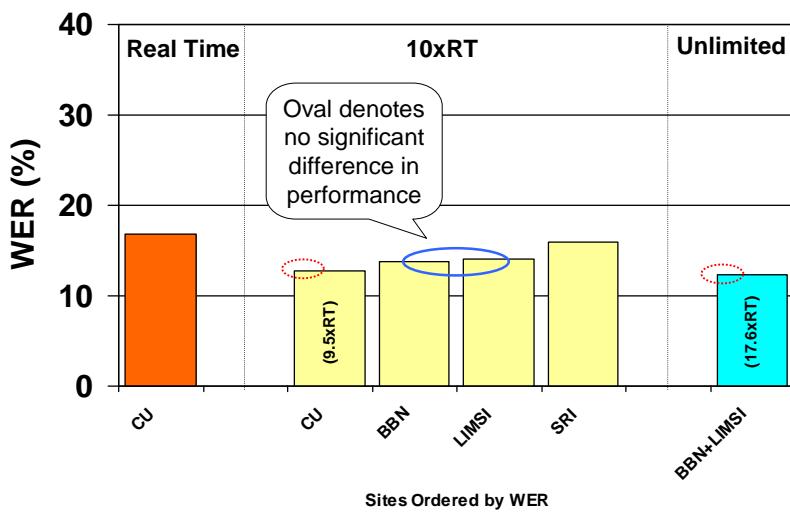
English BN WER by Gender Current Test Set



English BN WER vs. Processing Speed Current Test Set

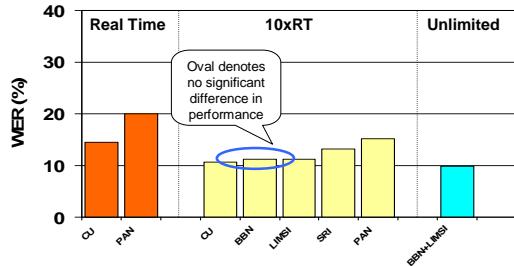


English BN STT Results Progress Test Set

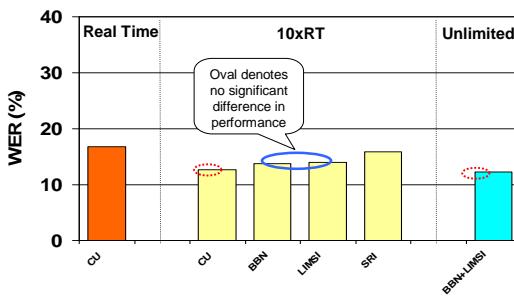


English BN STT Results

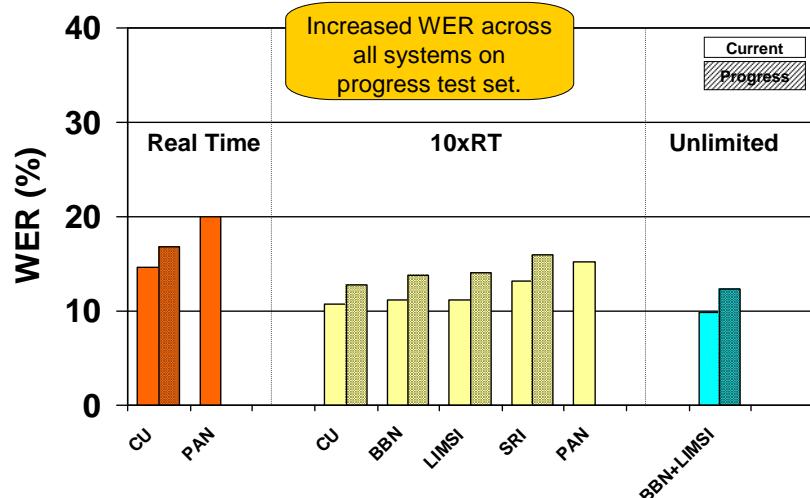
Current



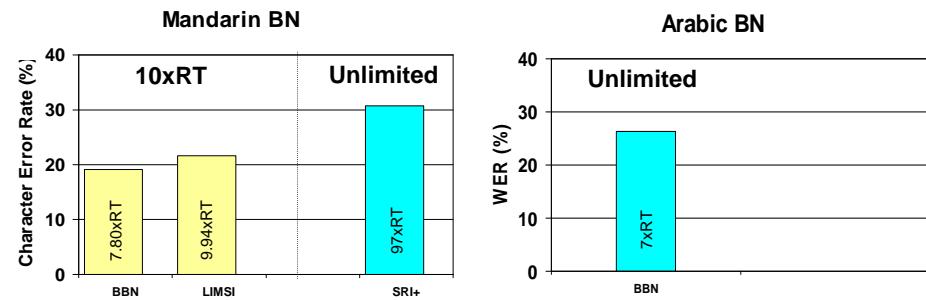
Progress



English BN WER Current vs. Progress



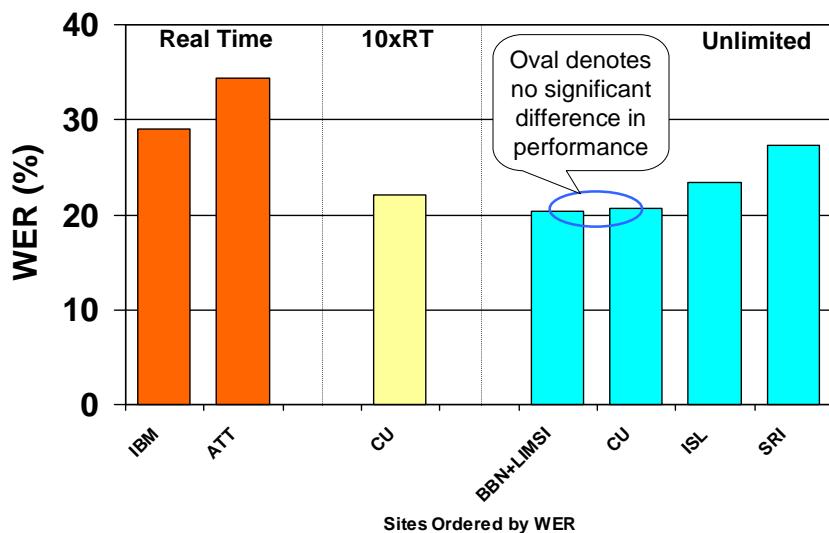
Non-English BN STT Results Current Test Set



CTS STT Results RT-03 Systems

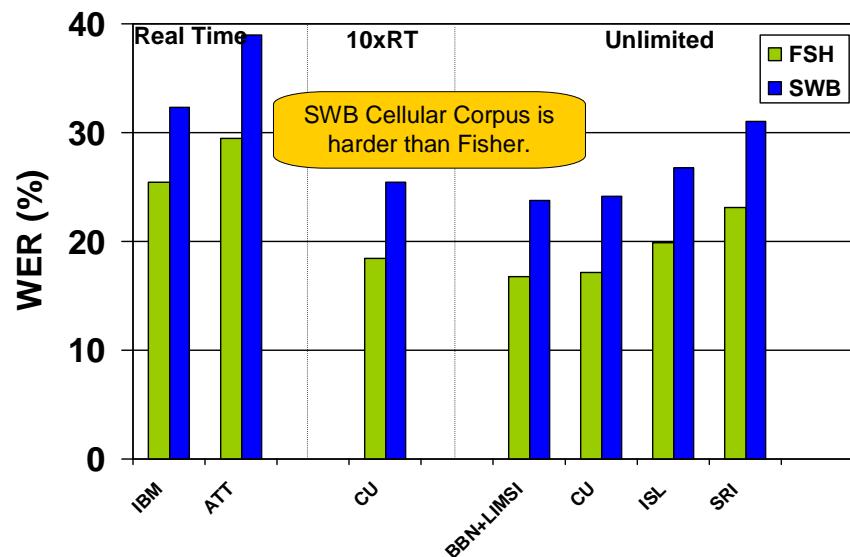
English CTS STT Results

Current Test Set



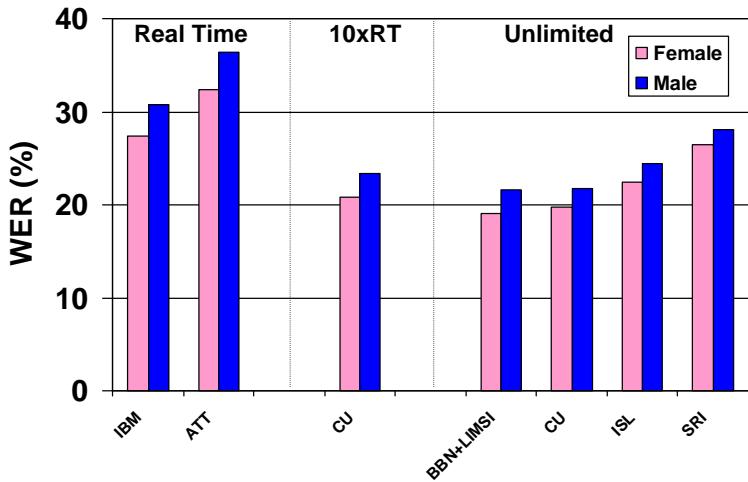
English CTS WER by Corpus

Current Test Set



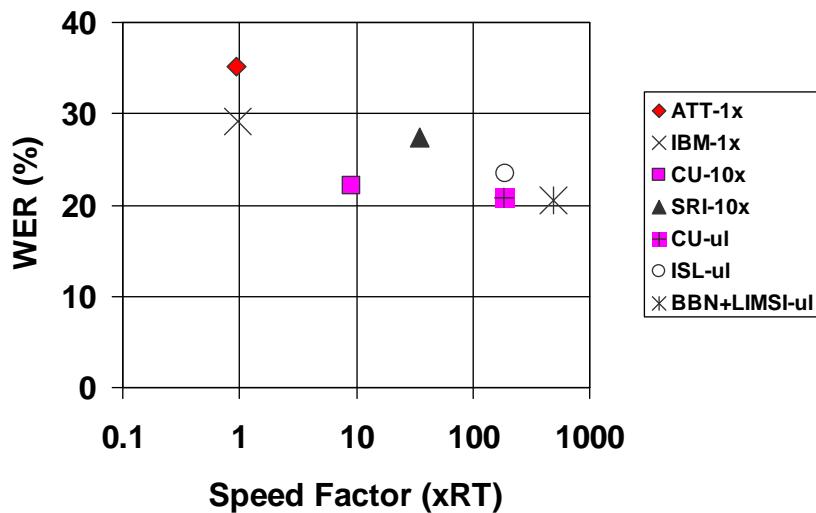
English CTS WER by Gender

Current Test Set



English CTS WER vs. Processing Speed

Current Test Set



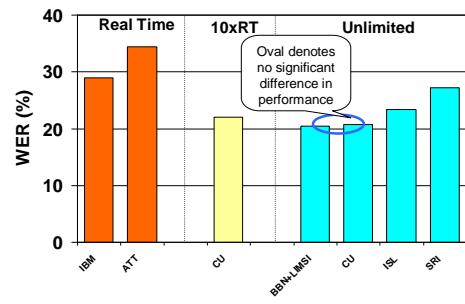
English CTS STT Results

Progress Test Set

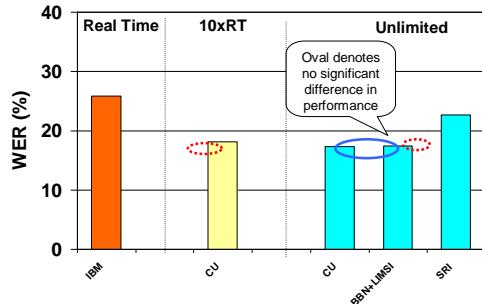


English CTS STT Results

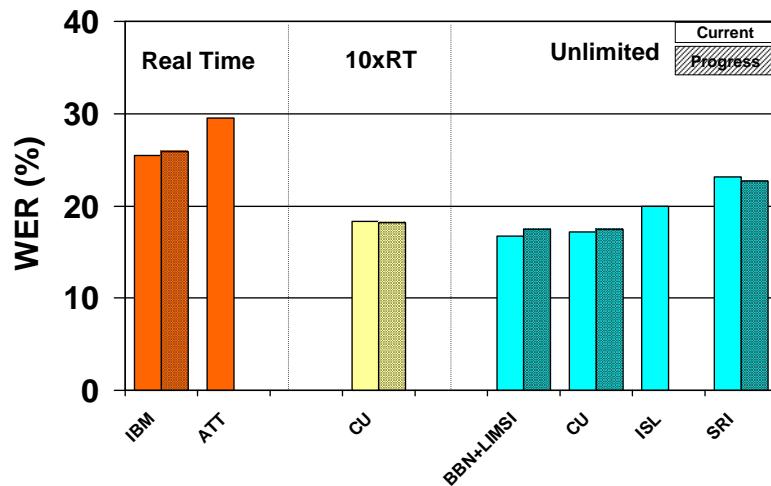
Current



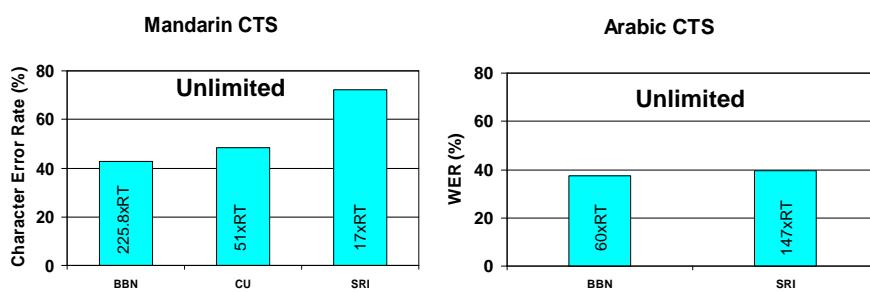
Progress

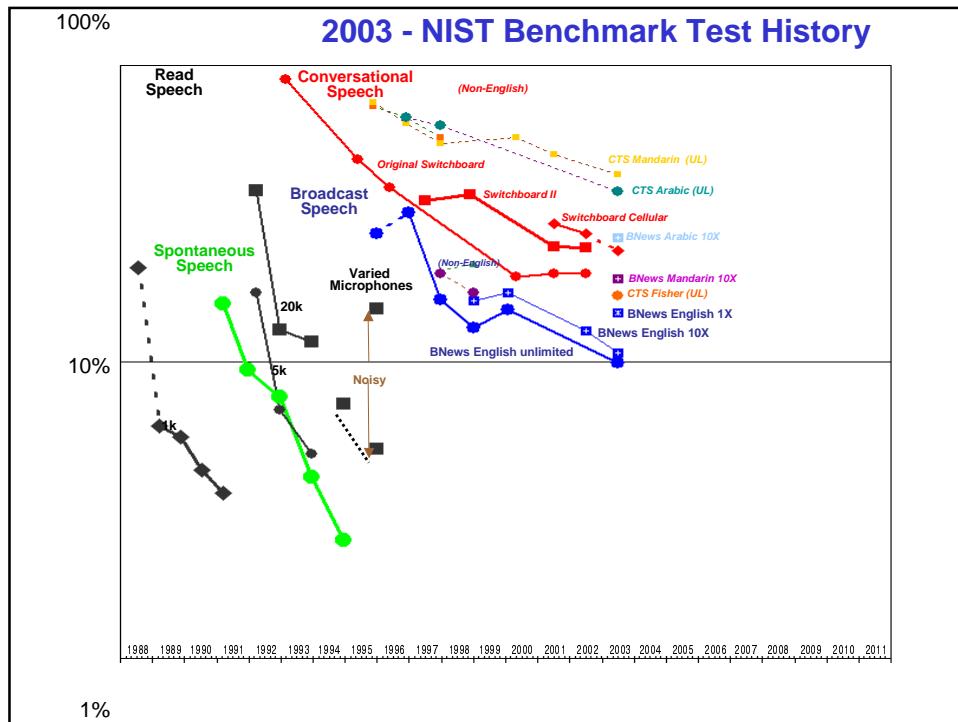


English CTS Fisher WER Current vs. Progress



Non-English CTS STT Results Current Test Set



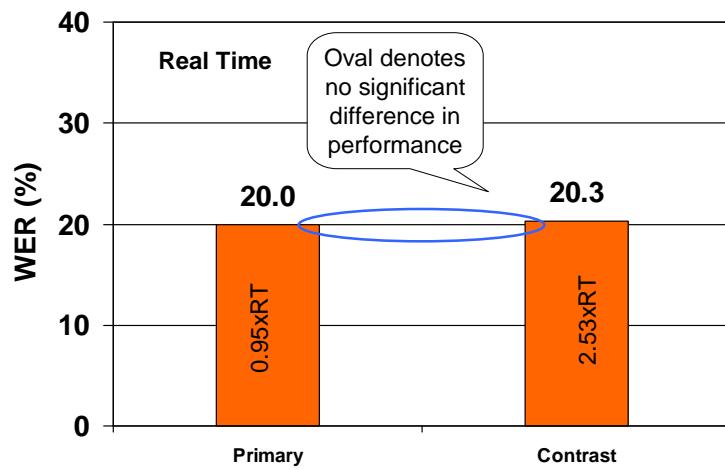


Conclusions

- Large number of test conditions but not a lot of participation in all of these conditions
 - Hard to make comparison
- Non-English WER is higher than English WER
- Significant improvement in current systems over last year's systems
- Future analysis
 - WER difference between Fisher and SWBD-Cell

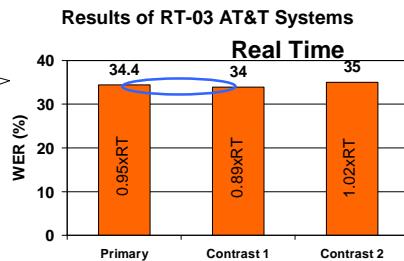
Appendix

Panasonic Primary vs. Contrastive Systems English BN STT Results Current Test Set



Primary vs. Contrastive Systems English CTS STT Results Current Test Set

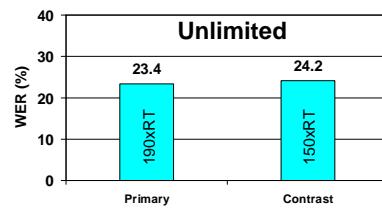
Oval denotes no significant difference in performance



Results of RT-03 SRI Systems



Results of RT-03 ISL Systems



English BN STT Results Current Test Set

Site	Test Set	System	Primary	Debug	Late	Language	Domain	Condition	WER(%)	NCE	Site SF
CU	eval03	cuhtk_1	yes	no	no	eng	bnews	stt1x	14.6	0.002	0.88
PAN	eval03	png	yes	no	no	eng	bnews	stt1x	20.0	-0.073	0.95
PAN	eval03	pnglia	no	no	no	eng	bnews	stt1x	20.3	-0.009	2.53
BBN	eval03	primary	yes	no	no	eng	bnews	stt10x	11.2	0.333	7.50
CU	eval03	cuhtk_1	yes	no	no	eng	bnews	stt10x	10.7	0.412	9.10
PAN	eval03	png	yes	no	no	eng	bnews	stt10x	15.2	0.002	9.20
LIMSI	eval03	primary	yes	no	no	eng	bnews	stt10x	11.2	0.379	9.50
SRI	eval03	sri1	yes	no	no	eng	bnews	stt10x	13.2	0.190	10.08
BBN+LIMSI	eval03	primary	yes	no	no	eng	bnews	sttul	9.9	0.203	17.10

STT English Summary Results													
STT English BNews %WER (Primary only)													
Site	Progress Set			Eval '03 Current Set			Site	Progress Set			Eval '03 Current Set		
	1X RT03	10X RT03	UL RT03	1X RT03	10X RT03	UL RT03		1X RT03	10X RT03	UL RT03	1X RT03	10X RT03	UL RT03
BBN			13.8								11.2		
BBN+LIMSI				12.3								9.9	
CU	16.8	12.7						14.6	10.7				
LIMSI		14.0								11.2			
PAN					20.0				15.2				
SRI			15.9						13.2				
STT English CTS %WER (Primary Only)													
Site	Progress Set			Eval '03 Current Set			Site	Progress Set			Eval '03 Current Set		
	1X RT03	10X RT03	UL RT03	1X RT03	10X RT03	UL RT03		34.4					
ATT													
BBN													
BBN+LIMSI				17.5							20.4		
CU		18.2	17.4						22.1		20.7		
IBM	25.9				29.0								
ISL										23.4			
LIMSI				22.7							27.3		
STT English CTS %WER (Primary Only)													
Site	Eval '03 Current Set (FSH)			Eval '03 Current Set (SWB)			Site	Eval '03 Current Set (FSH)			Eval '03 Current Set (SWB)		
	1X RT03	10X RT03	UL RT03	1X RT03	10X RT03	UL RT03		29.5			39.0		
ATT		29.5											
BBN													
BBN+LIMSI				16.7							23.8		
CU		18.4	17.1						25.5		24.1		
IBM	25.4				32.4						26.7		
ISL				19.9									
LIMSI					23.1						31.1		
SRI													

English BN STT Results											
Progress Test Set											
Site	Test Set	System	Primary	Debug	Late	Language	Domain	Condition	WER(%)	NCE	Site SF
CU	prog	cuhtk1	yes	no	no	eng	bnews	stt1x	16.8	0.001	0.97
CU	prog	cuhtk1	yes	no	no	eng	bnews	stt10x	12.7	0.387	9.45
SRI	prog	sri1	yes	no	no	eng	bnews	stt10x	15.9	0.213	10.08
BBN	prog	prim	yes	no	no	eng	bnews	stt10x	13.8	0.321	7.50
LIMSI	prog	primary	yes	no	no	eng	bnews	stt10x	14.0	0.354	9.80
BBN+LIMSI	prog	primary	yes	no	no	eng	bnews	sttul	12.3	0.192	17.60

Mandarin BN STT Results

Current Test Set

Site	Test Set	System	Primary	Debug	Late	Language	Domain	Condition	WER(%)	NCE	Site SF
BBN	eval03	prim	yes	no	no	mand	bnews	stt10x	19.1	0.397	7.80
LIMSI	eval03	primarylate	yes	no	yes	mand	bnews	stt10x	21.7	0.402	9.94
SRI+	eval03	decipherdebug	yes	yes	no	mand	bnews	stt1	30.8	-19.143	97.00

Arabic BN STT Results

Current Test Set

Site	Test Set	System	Primary	Debug	Late	Language	Domain	Condition	WER(%)	NCE	Site SF
BBN	eval03	primary	yes	no	no	arab	bnews	stt10x	26.3	-0.972	7.00

English CTS STT Results

Current Test Set

Site	Test Set	System	Primary	Debug	Late	Language	Domain	Condition	WER(%)	NCE	Site SF
ATT	eval03	trillseg	no	no	no	eng	cts	stt1x	35.0	-0.179	0.89
ATT	eval03	triattseg	yes	no	no	eng	cts	stt1x	34.4	-0.173	0.95
IBM	eval03	speedy	yes	no	no	eng	cts	stt1x	29.0	0.016	0.97
ATT	eval03	pentallseg	no	no	no	eng	cts	stt1x	34.0	-0.160	1.02
CU	eval03	cuhtk_1late	yes	no	yes	eng	cts	stt10x	22.1	0.318	9.21
SRI	eval03	sri1	yes	no	no	eng	cts	sttul	27.3	0.198	35.00
SRI	eval03	sri2late	no	no	yes	eng	cts	sttul	25.5	0.179	59.00
ISL	eval03	janus2	no	no	no	eng	cts	sttul	24.2	-5.176	150.00
CU	eval03	cuhtk_1	yes	no	no	eng	cts	sttul	20.7	0.318	186.85
ISL	eval03	janus	yes	no	no	eng	cts	sttul	23.4	0.000	190.00
BBN+LIMSI	eval03	primary	yes	no	no	eng	cts	sttul	20.4	0.209	486.00

English CTS STT Results

Progress Test Set

Site	Test Set	System	Primary	Debug	Late	Language	Domain	Condition	WER(%)	NCE	Site SF
IBM	prog	speedy	yes	no	no	eng	cts	stt1x	25.9	0.017	0.97
CU	prog	cuhtk1	yes	no	no	eng	cts	stt10x	18.2	0.312	8.89
SRI	prog	sri2	yes	no	no	eng	cts	sttul	22.7	-28.614	49.00
BBN+LIMSI	prog	primary	yes	no	no	eng	cts	sttul	17.5	0.195	393.60
CU	prog	cuhtk1	yes	no	no	eng	cts	sttul	17.4	0.303	191.34

Mandarin CTS STT Results

Current Test Set

Site	Test Set	System	Primary	Debug	Late	Language	Domain	Condition	WER(%)	NCE	Site SF
SRI	eval03	sri1debug_2	yes	yes	no	mand	cts	sttul	72.3	na	62.00
BBN	eval03	primary	yes	no	no	mand	cts	sttul	42.7	0.149	255.80
CU	eval03	cuhtk1	yes	no	no	mand	cts	sttul	48.6	0.190	51.90

Arabic CTS STT Results

Current Test Set

Site	Test Set	System	Primary	Debug	Late	Language	Domain	Condition	WER(%)	NCE	Site SF
SRI	eval03	sri1	yes	yes	no	arab	cts	sttul	39.7	0.058	60.00
BBN	eval03	primary	yes	no	no	arab	cts	sttul	37.5	0.230	147.00